# Open Data in the Humanities: Data Sharing and Publication for Triadic Co-Creation

***Asanobu Kitamoto***[1,2]*

[1] *Center for Open Data in the Humanities (CODH), Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan*
[2] *National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan*
Email: kitamoto@nii.ac.jp

***Summary.*** Center for Open Data in the Humanities (CODH) was established in April 2017 with the aim of developing data science research in the humanities. CODH has the concept of "triadic co-creation" in which three types of stakeholders, namely humanities scholars, machines (computer scientists), and citizens, collaborate together to advance data creation, analysis and utilization. In particular, we focus on the redefinition of the division of roles between humans and machines by providing large-scale open training data for machine learning (artificial intelligence). We already released several open datasets, such as books, characters, and recipes that are collaboratively created with National Institute of Japanese Literature under NIJL-NW project, and other datasets such as photographs, maps, geographic information, and magazines with other collaborators. We also work on data publication in the humanities, using IIIF (International Image Interoperability Framework) as one of platforms for sharing data as materials for humanities research.

***Keywords.*** Humanities, Open data, Triadic co-creation, data publication, IIIF (International Image Interoperability Framework).

## 1. Introduction

Center for Open Data in the Humanities (CODH) is promoting research and support activities aimed at opening data and promoting data-driven collaboration in the humanities discipline. In the humanities research community, however, data science approach based on large-scale open data is still premature and we cannot expect that the usage of open data increases without effort. CODH will therefore develop and release databases and tools by adopting the methodology of Digital Humanities (Computers and Humanities), which shows a rapid growth in the global research community, and also hold seminars and tutorials to promote the utilization of research resources.

## 2. Triadic Co-Creation

CODH has the concept of "triadic co-creation" in which three types of stakeholders, namely humanities scholars, machines (computer scientists), and citizens, collaborate together to advance data creation, analysis and utilization. In particular, we focus on the redefinition of the division of roles between humans and machines by providing large-scale open training data for machine learning (artificial intelligence).

Thanks to the development of machine learning technology in recent years, some work can be transferred from humans to machines, but in order to ask machines to work, we need to provide training data in the first place. The lack of such open training data leads to the delay of introducing machines to humanities research, so CODH will work on constructing infrastructure for open training data on which researchers and citizens can learn about and co-create open training data.

## 3. Open Data

We already released several open datasets, such as books, characters, and recipes that are collaboratively created under "Project to Build an

International Collaborative Research Network for Pre-modern Japanese Texts" (NIJL-NW Project), promoted mainly by National Institute of Japanese Literature [1].

First, "Dataset of Pre-Modern Japanese Text (PMJT)" provides image data of 701 pre-modern books in a downloadable format. By assigning DOI (Digital Object Identifier) to each book, image data can be uniquely identified even when there are multiple books with the same title.

Second, "Dataset of PMJT Character Shapes" contains 3,999 character types and 403,242 characters of old Japanese characters (called Kuzushi-ji). The dataset can be used not only as learning material for humans but also as training data for machines to develop optical character recognition (OCR) software. These datasets are also used for our "Kuzushi-ji Challenge!" campaign to involve experts and citizens to develop best artificial intelligence software that recognizes old Japanese characters.

Third, "Dataset of Edo Cooking Recipes" is the collection of recipes from a cooking book "Tamago Hyakuchin," which contains more than 100 recipes about eggs. We translated the original recipes to modern Japanese and structured for citizen cookers. Moreover, those recipe data was released in Japan's largest recipe service "Cookpad," which triggered unexpected excitement among citizens [2].

## 4. Data Sharing and Publication

Sharing data as open data is one form of scholarly publication in the age of open science. To find the best practice of data publication in the humanities, we are now focusing on IIIF (International Image Interoperability Framework) as infrastructure for image-related projects.

IIIF has recently seen a rapid growth of adoption as interoperable, high-resolution image delivery from museums and libraries around the world, and CODH is one of them to use IIIF for Pre-modern Japanese Text, and Digital Silk Road Digital Archive of Toyo Bunko Rare Books. While being widely adopted, IIIF is still a premature specification without important functions for humanities research.

CODH focuses on a use case of collecting interesting images from IIIF contents all over the world. We proposed a new specification called Curation API and developed a reference implementation called IIIF Curation Viewer, applied to PMJT curation and IIIF global curation.

Cropping images and collecting them under a theme is the basic task in the humanities research, and sharing them is one form of data publication having value as the material of subsequent research. We are now studying the potential of this approach in the field of art history.

## 5. Conclusion

CODH has other types of open data such as photographs, maps, geographic information and magazines. To organize and utilize those datasets, CODH also develops web and mobile applications for researchers and citizens. To publicize and disseminate open data in the humanities, various programs such as seminars, tutorials and courses should be helpful. Moreover, wider involvement of researchers and citizens into data creation and analysis activities, or citizen science, is another relevant challenge. CODH's activities are naturally across disciplines, and beyond disciplines in the sense of trans-disciplinary science.

## References


1. Kitamoto, A., Yamamoto, K., "High-throughput Collation Workflow for the Digital Critique of Old Japanese Books Using Computer Vision Techniques". *Sixth Annual Conference of the Japanese Association for Digital Humanities (JADH2016),* 2016
2. Kitamoto, A., "FAIRness for Citizens: Workflow and Platform for Open Data with a Case Study on Edo Cooking Recipes". *Seventh Annual Conference of the Japanese Association for Digital Humanities (JADH2017),* 14-16, 2017